DOCUMENT RESUME

ED 416 070 SE 061 058

AUTHOR Miller-Whitehead, Marie
TITLE Tennessee TCAP Science Scale Scores 1990-1997: Implications for Continuous Improvement and Educational Reform.
PUB DATE 1997-12-24
NOTE 16p.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Educational Change; Elementary Secondary Education; Science Curriculum; Science Education; *Scores; *Standardized Tests; Tables (Data); *Test Score Decline
IDENTIFIERS Tennessee

ABSTRACT

The goal of this research was to determine if standardized science test scores in Tennessee show evidence of continuous improvement in student achievement. The data examined as part of this study display evidence of a performance dip in Grade 4 but show an overall increase in scores across grade levels. A discussion of Tennessee's value added evaluation system and the implications that the analysis of this data can have on funding is included. The limitations of traditional means of evaluation in assessing a student's science learning are also discussed. Data displayed in table format pertain to minimum and maximum science scale scores, mean science scale scores, five-year mean science scale scores, science scale score descriptives, and an analysis of variance summary. Contains 30 references. (DDR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 416 070

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
M. Miller-Whitehead
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Tennessee TCAP science scale scores 1990 - 1997: Implications for continuous improvement and educational reform

Marie Miller-Whitehead
Education Consultant
December 24, 1997

SE041058

BEST COPY AVAILABLE

INTRODUCTION

Barely a week passes without some commentary, be it published in a scholarly journal or in the popular press, that addresses the issue of educational accountability and school reform. One of the most problematic issues is that of the negative effects of accountability and mandated testing on school reform initiatives, in particular the issue of multiple-choice testing versus performance testing and the relative merits of testing at all.

Critics of multiple-choice testing (Herman, Abedi, & Golan, 1994; Ligon & Wilkinson, 1985; Madaus, 1993; Perrone, 1991; Shepard, L. A., 1990) have, indeed, raised many valuable questions that deserve to be answered even if those answers must be qualified or stop short of being definitive. Is there evidence that multiple-choice testing suppresses hands-on learning or experiential learning? Does it suppress creativity in the classroom by encouraging a teach-to-the-test mentality? Does test preparation have a negative influence on innovation in the curriculum, or is it a useful tool for teachers and administrators who seek feedback to implement process improvement (Porter, 1983)? These questions most assuredly address well-traveled ground, but what has been lacking is empirical evidence that would support one position or the other. Most commentaries on the matter are either anecdotal or qualitative case studies, which is not to imply that the arguments are without merit. Nevertheless, it would seem within the reach of educators who have mandated multiple choice testing as a component of their accountability system to support opinions either way with quantifiable data in addition to the qualitative studies which

have been conducted.

Educational leaders should not only search for but demand empirical, quantifiable data that supports site-based and system-wide continuous improvement. To that end a variety of longitudinal studies have been undertaken in Tennessee using the TVAAS data set (Achilles, 1996; Achilles, Zaharias, & Nye, 1995; Finn & Achilles, 1990; Nye, 1993; Nye, 1992). The Tennessee Value-Added Assessment System has been in place and has been used as the statewide vehicle for computing and disseminating (with the State Testing and Evaluation Center) value-added gain scores from CTBS/4 test score data since 1992, with pilot testing and phase-in since the late 1980s. For this reason, educational administrators and policymakers have at hand a stable set of statistical data to aid in the decision-making process.

STATEMENT OF THE PROBLEM

How can this data be helpful to educators? The question to be answered is, "Is there evidence of continuous improvement in student achievement on the CTBS/4 science test from 1990 to 1997? An examination of Tennessee scale science scores provides that evidence and support for the findings of other researchers who seek to determine the effects of new programs and curriculum on student achievement. Eastwood (1993; Eastwood & Louis, 1992) documented the "performance dip" in studies of curricular changes, finding that during the learning curve that for both students and teachers that takes place after the implementation of a change it is not unusual to see a drop in overall student performance. This drop in student

achievement scores can be disconcerting, not to say discouraging, to policymakers unless they realize that this is a result which can be expected and which will correct itself after professional development and training of all personnel, faculty, and staff who are charged with the implementation of change.

The Tennessee science score data set displays evidence of the performance dip just as might have been predicted from the Eastwood study. While the TVAAS value-added reports are based on three years of data, the value of using a longitudinal data set can be seen by a close examination of the statewide data for five year periods, beginning in 1990 to the present time. With the proposal for and publicity given to the value-added assessment system in 1990, many systems began to plan for what would become the legislature-mandated accountability for student achievement beginning in 1992 with the EIA, or Education Improvement Act of Tennessee. Many schools implemented school improvement initiatives, changed their curriculum, or otherwise prepared for the upcoming accountability law beginning in 1990 or 1991 when the state legislature supported pilot studies using the CTBS/4 test and the value-added assessment system. According the the data, in the period between 1990 and 1996, only 1991 exhibited a drop in student science achievement scale scores (mean aggregate science scale score grades 2-8, 1991 = 721.42). For the six year period of 1992 - 1997 Tennessee statewide aggregate scores improved each year until 1997 (Table 2).

RESULTS

There is much encouraging information in these results, particularly in light of Tennessee's state mandate to provide equal funding to school systems to promote equality in education, but the picture is not entirely rosy. For example, while the overall aggregate mean scale scores have risen (Table 2), the maximum scores, denoting performance of students and schools at the upper end of the spectrum, have been uneven with the highest maximum scores occuring in year 1993 (max mean science scale score = 801.2) and, for grade eight, in year 1990 (max mean science scale score = 807.3). The minimum scores have also been uneven, ranging from 621.5 in 1997 to a high of 631.5 in 1995.

Examining the aggregate mean scale scores by grade level over five year periods (Table 3), mean scale scores have shown an increase for every grade level except for grade four, where the mean scores for the period 1993 - 1997 were lower than for the five year period from 1992 - 1996. These results indeed point to Tennessee's overall improvement in student achievement as measured by CTBS/4 science tests. It would appear that the funding changes have had an effect. Is the effect a significant change?

A basic course in statistics or knowledge of the central limit theorem and probability would lead most educators to assume that the effect of regression to the mean would be apparent, particularly in the minimum and maximum scores, with the minimum scores having a tendency to become higher and the maximum scores to

BEST COPY AVAILABLE

decline. However, any effect of regression to the mean might be expected to be offset by other factors, most notably the tendency of students to become "test wise" over time. This effect may be referred to as maturation. In other words, students who took the CTB test in 1990 had not had extensive practice in taking the CTB test, whereas students in 1997 had presumably taken one form or another of the test each year they had attended public school in Tennessee. Therefore, the expectation is that all other things being equal, student test taking skills should improve over time. Reliability coefficients for the CTB science tests administered by Tennessee range from .73 to .85, according to the Technical Manual. Those interested in the implications of those figures are referred to published reviews (Baker & Xu, 1995; Bock, Wolfe, & Fisher, 1996; Hopkins, 1992; Miller, 1992; Noble & Sawyer, 1992) and to standard texts in the areas of measurement and evaluation (McLean & Lockwood, 1996).

Now it is obvious that a certain amount of normal variation in the scores is to be expected. The question is whether the yearly variations were statistically significant, and if so, how. To that end ANOVA procedures were conducted on the science scale score data by grade level and by year. Results indicated that year was significant, with both 1996 and 1997 scale scores significantly better than 1993 scale scores ($F = 3.59$, $p < .05$, $R = .052$, Table 5). Even though the 1997 aggregate mean scale score was lower than that for 1996, the difference was not statistically significant. These are indeed positive results for Tennessee's progress in assuring continuous student improvement. The improvement is most assuredly of an

incremental nature, at least at the state level, and an examination of three years of data would not have revealed significant improvement; however, by looking at five consecutive years of data it is possible to detect the gradual but significant upward trend. These results are consistent with the NAEP findings (Blank, 1992; Campbell, 1996).

IMPLICATIONS

How do system-level educators best make use of this data? By comparing school and system level results with both national and Tennessee data, any school can track its progress, keeping in mind that it may expect to see not only random variation, but an occasional "performance dip," particularly the first year after major curricular changes. On the other hand, if the national trend and the state trend are both upward, then a system which finds its scores static over a period of several years should reasses curriculum offerings, alignment, and professional development of faculty and staff. Once again, a high performing school system with static scores may have determined that student performance indicators in such areas as science fairs, projects, and other innovative alternative assessments more than make up for standardized test scores which are consistently high but which are not showing improvement over time. In fact, a system which does not show improvement in standardized test scores and which does not provide adequate alternative methods for students to demonstrate excellence risks losing students to private schools. Given the high stakes nature of the Tennessee accountability system, it is crucial that administrators be able to communicate the

III. Document Availability Information

(Non-ERIC Source)

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents which cannot be made available through EDRS).

Publisher/Distributor:

Address:

Price Per Copy:

Quantity Price:

IV. Referral to Copyright/ Reproduction Rights Holder

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Return to the ERIC Database page.

strengths and weaknesses of the various accountability mechanisms in place in such a way that students, parents, and the community can make informed assessments of the educational quality of their schools. In the final analysis, much of educational choice and funding for public education is political, as Dorn (1998) in his most recent analysis of accountablity mechanisms has pointed out.

Turning to anecdotal evidence and first-hand knowledge as an observor of the teaching and learning taking place in the sciences, it would be very difficult to take the position that multiple-choice testing has discouraged innovation and creativity in Tennessee's classrooms. One has only to look at the wide variety of hands-on projects posted on WWW sites, visit classrooms, and speak with parents and teachers. These projects frequently generate enormous amounts of excitement, interest, and positive publicity for students, schools, and communities, and, in fact it is often through such projects that the "breakthrough" types of improvement (as opposed to incremental improvement) are demonstrated. It is nevertheless many times quite difficult to assess from such projects whether students have been exposed to the rich and comprehensive range of the curriculum without the support of some kind of standardized test.

BEST COPY AVAILABLE

Table 1

Minimum and Maximum Science Scale Scores for 1993 - 1997

Mean Score YEAR	Minimum	Maximum
93	627.9	801.2
94	625.7	792.5
95	631.5	797.8
96	628.2	799.0
97	621.5	793.1

NOTE. These are aggregate science scale scores for grades two through eight for 956 schools

Table 2

Mean Science Scale Scores for Grades 2 - 8 by Year

1993		1994		1995		1996		1997	
M	N	*M*	N	*M*	N	*M*	N	*M*	N
723.76	956	724.48	956	726.28	957	728.56	958	728.28	959

Table 3

5 Year Mean Science Scale scores for 1990 - 1997 by Grade Level

Grade	2	3	4	5	6	7	8
1990-1994	667.51	690.96	713.44	728.49	739.64	754.89	766.82
1992-1996	668.56	692.55	716.64	729.57	742.37	759.03	771.38
1993-1997	669.39 N=690	693.89 N=690	716.29 N=690	730.55 N=690	744.99 N=686	759.23 N=671	771.99 N=669

Table 4

Science Scale Score Descriptives by Grade Level and by Year, 1993-1997

	N	M	min	max	variance	SD
SS93.2	138	662.57	627.90	692.90	157.98	12.57
SS93.3	138	686.48	653.50	717.40	119.89	10.95
SS93.4	138	716.46	681.60	741.40	119.28	10.92
SS93.5	138	726.97	699.60	751.20	72.34	8.51
SS93.6	138	746.42	705.70	775.60	106.53	10.32
SS93.7	138	754.55	729.70	779.00	61.99	7.87
SS93.8	138	770.67	747.80	794.60	54.75	7.40
SS94.2	138	674.56	625.20	714.60	166.74	12.91
SS94.3	138	698.61	650.10	732.50	162.42	12.75
SS94.4	138	715.85	682.30	743.60	95.15	9.76
SS94.5	138	733.48	698.90	754.30	87.22	9.34
SS94.6	137	734.98	698.20	756.50	79.74	8.93
SS94.7	134	753.05	720.80	784.30	73.07	8.55
SS94.8	133	765.08	745.60	787.40	60.60	7.79
SS95.2	138	668.99	631.70	702.40	166.09	12.89
SS95.3	138	691.48	644.20	728.10	139.28	11.80
SS95.4	138	715.38	671.40	743.50	112.09	10.59
SS95.5	138	727.88	696.70	771.60	103.30	10.16
SS95.6	137	747.45	722.50	784.40	121.15	11.01
SS95.7	134	764.37	732.50	788.90	81.62	9.03
SS95.8	134	772.34	743.40	796.60	63.06	7.94
SS96.2	138	675.51	629.80	713.10	219.84	14.83
SS96.3	138	699.24	642.60	729.90	196.50	14.02
SS96.4	138	717.88	666.70	747.00	122.47	11.07
SS96.5	138	731.47	694.20	759.10	109.71	10.47

Tennessee TCAP science scale scores 1990 - 1997

	N	M	min	max	variance	SD
SS96.6	137	744.91	713.50	779.40	105.30	10.26
SS96.7	135	760.96	730.30	788.60	94.04	9.70
SS96.8	134	774.49	744.70	798.30	73.15	8.55
SS97.2	138	673.24	621.50	714.90	203.35	14.26
SS97.3	138	698.87	650.60	737.60	154.41	12.43
SS97.4	138	716.08	683.10	742.00	119.35	10.92
SS97.5	138	733.02	689.00	766.00	118.41	10.88
SS97.6	137	747.80	710.50	778.90	113.52	10.65
SS97.7	135	758.06	726.40	786.40	80.14	8.95
SS97.8	135	769.85	741.80	793.10	61.58	7.85

Table 5

Analysis of Variance Summary Table for Year Effect, Years 1993-1997

Source	SS	df	MS	F	η^2
Year	18006.81	4	4501.70	3.59*	0.003
Error	5992796	4781	1253.46		
Total	6010802	4785			

*p< .05

LIST OF REFERENCES

Achilles, C. M. (1996). Students achieve more in smaller classes. Educational Leadership, 53(5), 76-77.

Achilles, C. M., Zaharias, J. B., & Nye, B. A. (1995). Analysis of Policy Application of Experimental Results: Project Challenge. (ERIC Document Reproduction Service No. ED 393 151)

Baker, A. P., & Xu, D. (1995). The Measure of Education: A Review of the Tennessee Value Added Assessment System. Tennessee State Comptroller of the Treasury, Nashville Office of Educational Accountability.

Blank, R. K., & Engler, D. (1992). Has science and mathematics education improved since "A Nation at Risk": Trends in course enrollments, qualified teachers, and student achievement. Science and Mathematics Indicators Project. Council of Chief State School Officers, Washington, DC. State Education Assessment Center.

Bock, R. D., Wolfe, R., & Fisher, T. H. (1996). A Review and Analysis of the Tennessee Value-Added Assessment System. Nashville, TN: Office of Education Accountability.

Bruschi, B. A., & Anderson, B. T. (1994). Gender and Ethnic Differences in Science Achievement of Nine-, Thirteen-, and Seventeen-Year-Old Students. (ERIC Document Reproduction Service No. ED 382 751)

Campbell, J. R., et al. (1996). NAEP 1994 Trends in Academic Progress. Report in Brief. Educational Testing Service, Princeton, NJ. Center for the Assessment of Educational Progress., National Assessment of Educational Progress, Princeton, NJ.

Dorn, S. (1998). The political legacy of school accountability systems. Education Policy Analysis Archives, [On-line serial] 6(1). Available WWW http://olam.ed.asu.edu/epaa/

Eastwood, K. W. (1993). The performance dip. American School Board Journal, 180(2), 47-48.

Eastwood, K. W., & Louis, K. S. (1992). Restructuring that lasts: Managing the performance dip. Journal of School Leadership, 2(2), 212-24.

BEST COPY AVAILABLE

Finn, J. D., & Achilles, C. M. (1990). Answers and Questions about Class Size: A Statewide Experiment. American Educational Research Journal, 27(3), 557-77.

Herman, J. L., Abedi, J., & Golan, S. (1994). Assessing the effects of standardized testing on schools. Educational and Psychological Measurement, 54(2), 471- 482.

Hopkins, K. D. (1992). Review of the Comprehensive Test of Basic Skills, Fourth Edition. In J. J. Kramer & J. C. Conoley (Eds.), The eleventh mental measurements yearbook. Lincoln: University of Nebraska, Buros Institute of Mental Measurement.

Johnson, E. G., et al. (1992). The NAEP 1990 Technical Report. National Assessment of Educational Progress, Princeton, NJ.

Ligon, G., & Wilkinson, D. (1985). The average achievement test score: A demogogue statistic. Paper presented at the 69th annual meeting of the American Educational Research Association, Chicago, IL.

Madaus, G. F. (1993). The distortion of teaching and testing: High stakes testing and instruction. Peabody Journal of Education, (2), 28-45.

McLean, J. E., & Lockwood, R. (1996). Why we assess students - and how: The competing measures of student performance. Thousand Oaks, CA: Corwin Press, Inc.

Miller, M. D. (1992). Review of the Comprehensive Test of Basic Skills, Fourth Edition. In J. J. Kramer & J. C. Conoley (Eds.), The eleventh mental measurements yearbook. Lincoln: University of Nebraska, Buros Institute of Mental Measurement.

Miller-Whitehead, M. (1997). An analysis of science scale scores for grades 2-8 in Tennessee for 1990 -1994. Unpublished manuscript.

Noble, J., & Sawyer, R. (1992). A Comparison of Two Approaches for Measuring Educational Growth from CTBS and P-ACT+ Scores. (ERIC Document Reproduction Service No. ED 346 163)

Nye, B. (1993). Some Questions and Answers about Multiage Grouping. (ERIC Document Reproduction Service No. ED 384 998)

Nye, B. A., et al. (1992). Smaller Classes Really Are Better. American School

Board Journal, 179(5), 31-33.

O'Sullivan, C. (1995). The Cost of Performance Assessment in Science: The NAEP Perspective. Educational Testing Service, Princeton, N.J. (ERIC Document Reproduction Service No. ED 384 638)

Perrone, V. (1991). The abuses of standardized testing. Childhood Education, 67(3), 131-142.

Porter, A. (1983). The role of testing in effective schools. American Education, 19(1), 25 - 28.

Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. Educational Policy Analysis Archives, [On-line serial], 3(6). Available WWW: http://olam.ed.asu.edu/epaa/.

Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement (R11-0435-02-001-97). Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Shepard, L. A. (1990). Inflated test score gains: Is it old norms or teaching to the test? Educational Measurement: Issues and Practices, 9(3), 15 - 22.

Underwood, S., & Lumsden, L. S. (1994). Class Size. ERIC Clearinghouse on Educational Management, Eugene, Oreg.; National Association of Elementary School Principals, Alexandria, VA. Office of Educational Research and Improvement (ED), Washington, DC. (ERIC Document Reproduction Service No. ED 377 548)

Yepes-Baraya, M. (1995). Task Analysis of Science Performance Tasks and Items: Identifying Relevant Attributes. Educational Testing Service, Princeton, N.J. (ERIC Document Reproduction Service No. ED 388 676)